Impact of Shutdown Techniques for Energy-Efficient Cloud Data Centers

Issam Raïs[1], Anne-Cécile Orgerie[2], and Martin Quinson[3]

[1] Inria, LIP, Lyon, France. issam.rais@inria.fr
[2] CNRS, IRISA, Rennes, France. anne-cecile.orgerie@irisa.fr
[3] ENS Rennes, IRISA, Rennes, France. martin.quinson@ens-rennes.fr

Abstract. Electricity consumption is a worrying concern in current large-scale systems like datacenters and supercomputers. The consumption of a computing unit is not power-proportional: when the workload is low, the consumption is still high. Shutdown techniques have been developed to adapt the number of switched-on servers to the actual workload. However, datacenter operators are reluctant to adopt such approaches because of their potential impact on reactivity and hardware failures. In this article, we evaluate the potential gain of shutdown techniques by taking into account shutdown and boot up costs in time and energy. This evaluation is made on recent server architectures. We also determine if the knowledge of future is required for saving energy with such techniques. We present simulation results exploiting real traces collected on different infrastructures under various machine configurations with several shutdown policies, with and without workload prediction.

1 Introduction

In order to make data centers more energy-efficient, a wide variety of approaches have been proposed in the recent years, ranging from free cooling to low-power processors, and tackling wasted watts at each level of the data center [3]. While such an on/off approach has been extensively studied in literature, most infrastructure administrators still dare not use it in their datacenters. This situation is due to two factors: firstly, until very recently, servers were not designed to be switched off; secondly, switching off takes time and energy. So it is difficult for administrators to estimate their potential energy gains versus their potential loss of reactivity due to a too long booting time. Several solutions have been proposed to limit this possible performance impact, like keeping few nodes idle or using hibernation or standby modes to fasten the boot.

In this paper, we study different shutdown techniques for computing resources in data centers, like actual switching off and hibernation modes. Moreover, we estimate the impact of such techniques on the energy consumption, the reactivity of the platform and on the lifetime of the servers. Our validations combines real power measurements and real datacenter traces with simulation tools.

The main contributions of this paper consists in:

1. evaluating the impact of shutdown techniques (ie. switching off unused servers) on the energy consumption;
2. showing the impact of such shutdown techniques on disk lifetime and energy consumption with and without workload prediction algorithm;

The reminder of this paper is structured as follows. Section 2 presents the related work. The on/off energy model and the shutdown policies are introduced in Section 3. The experimental setup is provided in Section 4. The experimental validation is shown in Section 5. Finally, Section 6 concludes this work and presents the future directions.

2 Related work

Shutdown techniques require 1) the hardware ability to remotely switch on and off servers, and 2) energy-aware algorithms to timely employ such an ability. This section describes the state-of-the-art approaches for both features.

2.1 Suspend modes on Linux kernel

We focus on the Linux implementation of ACPI specification system power management. The available sleep states on the Linux kernel are:

- S0 or "Suspend to Idle" : freezing user space and putting all I/O devices into low-power states
- S1 or "Standby / Power-On Suspend" : same as S0 adding the fact that non boot CPUs are put in offline mode and all low-level systems functions are suspended during transitions into this state. The CPU retains power meaning operating state is lost, so the system easily starts up again where it left off
- S3 or "Suspend-to-RAM" : Everything in the system is put into low power state mode. System and device state is saved and kept in memory (RAM).
- S4 or "Suspend-to-disk" : Like S3, adding a final step of writing memory contents to disk.
- S5 or "System shutdown state" : Similar to S4, except that the OS doesn't save any context.

On the top of our knowledge, many datacenters servers do not implement or allow S3 (Suspend-to-RAM) sleep state, because of numerous errors when resuming (especially errors due to network connections with Myrinet or Ethernet protocols). Typically, only S0, S5(regular shutdown) are available for operational use.

2.2 Shutdown policies

The resource manager is responsible for deciding when to suspend and resume nodes. It takes decisions either based on pre-determined policy [8] or on workload

predictions [4]. In this paper, we study simple shutdown techniques, without combining them to scheduling algorithms in order to evaluate the impacts of such techniques without interfering with the workload of real platforms and with the users' expected performances.

The main disadvantage of shutdown policies resides in the energy and time losses that may occur when switching off and on takes longer than the actual idle period. The various suspend modes offer different performances concerning the time they need to switch between the On and Off states and the energy they consume while in Off state. The next section provides formalism for evaluating the impact of parameters for shutdown techniques.

3 Models

In this section, we describe the different models used by the shutdown policies we want to evaluate in order to determine when a node has to be switched off.

3.1 Energy efficiency time threshold

Switching on and off a server consumes time and energy, it is thus required to take these costs into account when deciding whether to switch off an idle server or not. In [5], the authors introduce T_s a time threshold such that if the server is idle for less than T_s, it should remain idle to save energy. Moreover, T_s needs to be greater than the time requited to switch off and on again a server in order for this threshold to be physically acceptable.

In order to compute T_s, all parameters described in it's definition [5] have to be known for each concerned server. These parameters can be acquired through a calibration measurement campaign. Then a shutdown policy is required to know when to switch off nodes. Indeed, as future is not known in the general case, it is difficult to determine for a given idle data center server if it will stay idle for more than T_s or not.

3.2 Studied shutdown policies

As the goal of this paper is to evaluate the impacts of on/off strategies rather than proposing new shutdown policies, we chose to lean on two ideal policies which will provide theoretical values about energy consumption.

Policy P1: knowing the future In this first policy, we consider that the future is completely known. Thus, dates and lengths of idle period are known for each server. This policy will give a theoretical lower bound for energy consumption with a perfect prediction algorithm.

Policy P2: aggressive shutdown The second policy does not consider the future and tries to switch off a server as soon as it is in idle state without any prediction attempt. Such an aggressive approach is expected to result in a higher energy consumption than Policy 1 because some idle periods may be lower than T_s. In such cases, switching off increases the energy consumption compared to

staying idle. This policy provides a simplified version of actual algorithms that wait for a given amount of time (usually greater than T_s) before switching off idle nodes.

These two policies depict a representative sample of typical shutdown policies deployed on real data centers. They will be compared in order to provide an evaluation of the potential impacts of such policies on energy consumption and nodes lifetime.

4 Experiment setup

In order to provide a fair comparison among policies P1 and P2, we simulate their behavior on real workload traces. The simulation tool is using real diversified calibration measurements. Simulations combine the workload traces and the energy calibration values to compare the two policies according to relevant metrics presented at the end of this section.

4.1 Workload traces

The utilized workload traces come from two kinds of data centers providing two different utilization scenarios which exhibit different workload patterns and utilization levels.

Operational Cloud platform: E-Biothon The E-Biothon platform is an experimental Cloud platform to help speed up and advance research in biology, health and environment [2]. It is based on four Blue Gene/P racks and a web portal that allow members of the bioinformatics community to easily launch their scientific applications. Overall, the platform offers 4096 4-cores nodes, reaching a peak power of 56 TeraFlop [2]. We obtained a workload trace for this platform covering from the 1st of January 2015 to the 1st of April 2016, so roughly 15 months of resource utilization.

Experimental testbed: Grid'5000 Grid'5000 is a large-scale and versatile testbed for experiment-driven research in all areas of computer science, with a focus on parallel and distributed computing including Cloud, HPC and Big Data [1]. For our evaluation, we took the workload trace of the Rennes site from the 1st of April 2010 to the 1st of April 2016, thus representing 6 years of resource utilization on this site. During this period, the weighted arithmetic mean of the number of nodes is 149.

4.2 Energy calibration

Grid'5000 provides management tools like kapower3, a utility that allows a user to have control on the power status of a reserved node[4], and, on some sites, it gives access to external wattmeters monitoring entire servers with a 0.125 Watts accuracy. This infrastructure is used for obtaining the energy calibration measurements required to compute T_s as described in Section 3.1.

[4] `https://www.grid5000.fr/mediawiki/index.php/Power_State_Manipulation_commands`

Table 1. Calibration nodes' characteristics and energy parameters for On-Off and Off-On sequences (average on 100 experimental measurements)

Features	Orion	Taurus	Paravance
Server model	Dell PowerEdge R720	Dell PowerEdge R720	Dell PowerEdge R630
CPU model	Intel Xeon E5-2630	Intel Xeon E5-2630	Intel Xeon E5-2630v3
Number of CPU	2	2	2
Cores per CPU	6	6	8
Memory (GB)	32	32	128
Storage (GB)	2 x 300 (HDD)	2 x 300 (HDD)	2 x 600 (HDD)
GPU	Nvidia Tesla M2075	-	-
Parameters	Orion	Taurus	Paravance
E_{OffOn} (Joules)	23,386	19,000	19,893
E_{OnOff} (Joules)	2,300	2,000	2,000
T_{OffOn} (seconds)	150	150	167.5
T_{OnOff} (seconds)	10	10	7.5
P_{idle} (Watts)	135	95	150
P_{off} (Watts)	18.5	8.5	4.5
T_s (seconds)	195	227	172

The results presented on the bottom part of Table 1 show values for regular shutdown, S5 mode (average of 100 run).

4.3 Evaluation metrics

In order to fairly compare the shutdown policies in the determined use cases, we define several evaluation metrics. In particular, for evaluating their energy impact, we compare the energy consumed with each policy against the energy used without any shutdown policy (ie. policy where the nodes stays idle and consumes P_{idle} Watts during periods without any work). This metric will indicate the potential energy savings with each policy.

We also provide the theoretical maximum energy savings if switching operations had a null cost (ie. zero energy, zero time for switching between on and off states). This provides an idea on how far the policies are from the theoretical ideal case and how much the costs related to switching operations are impacting the energy savings. The ideal case does not provide 100% energy gains compared to the idle case as switched off nodes consume energy ($P_{off} \neq 0$).

Finally, the results include the number of On-Off cycles per node for each workload in order to evaluate the impact of shutdown policies on the servers' lifetime. Indeed, one obstacle to the adoption of shutdown policies lies in the number of On-Off cycles imposed to the servers. In case of a too high number of cycles, it could damage the hardware parts like the hard disk drives (HDD). Typically, it is considered that hard drives can support a given amount of switching on and off during their lifetime. This parameter, known as *Contact Start/Stop Cycles* or *load/unload cycles* depending on the physical configuration of the hard drive head, is typically around 50,000 and 300,000 respectively for desktop HDD [6],

and around 600,000 for NAS HDD (Network-Attached Storage) which use only load/unload technology [7]. So, the number of On-Off cycles per node will be compared with these figures to determine whether the policy may alter or not the servers' lifetime.

5 Experiments: Simulation results based on actual hardware calibration

This section explores the simulation results of the shutdown policies with the various hardware calibrations and the workload traces described in Section 4. For every trace replay, the nodes are assumed to be homogeneous. Thus, every node of the trace is respecting the configuration of one of the calibrated nodes for each run.

5.1 Impacts of shutdown policies and prediction influence on energy consumption

We examine the case of current architectures based on the calibration made on the Grid'5000 nodes and described in Table 1.

Table 2 shows the percentage of energy that could be saved during idle periods with each policy compared to the energy consumed if nodes are never switched off. The last two columns present the average number of On-off cycles per node for the entire duration of the workload.

Table 2. Energy gains on idle periods and number of on-off cycles per node for current servers

	% Energy saved on idle periods			# On-Off cycles per node	
Calibration	P1	P2	Ideal	P1	P2
Grid'5000 trace, 6 years, 149 nodes on average					
Orion	85.87%	85.59%	86.29%	3,080	5,690
Taurus	90.56%	90.22%	91.05%	2,980	5,690
Paravance	96.66%	96.46%	97.00%	3,333	5,690
E-Biothon trace, 15 months, 4096 nodes					
Orion	85.18%	84.56%	86.29%	33	70
Taurus	89.83%	89.07%	91.05%	33	70
Paravance	96.03%	95.61%	97.00%	38	70

The results show that by turning off nodes, even when considering On-Off and Off-On costs, consequent energy gains can be made on real platforms. In the most unfavorable configuration (ie. Orion configuration), we can theoretically save up to 86% of the energy consumed while being in idle state. In the case of Grid'5000 trace, this percentage represents around 706,000 kWh for the 6 years. For the E-Biothon trace, we can also save up to 86% of the energy consumed in

the idle case, this represents 109,000 kWh for 15 months of loss to keep servers idle.

The number of On-Off cycles per node reaches at the maximum 5,690 for the 6-year Grid'5000 traces, far less than the 50,000 start/stop cycles typically allowed by HDD manufacturers [6,7]. This clearly states that even aggressive shutdown policies have no impact on disk lifetime.

It is worth noticing that significant energy gains can be performed for both traces even though they present completely different use cases. In particular, the E-Biothon trace comes from an operational bioinformatics supercomputer and although energy savings are smaller than for the Grid'5000 trace in comparison with the infrastructure size, they are still not negligible, representing around 73,680 kWh per year for the Orion case (most unfavorable case) with a basic shutdown policy like P2 (without prediction algorithm).

The energy saved with policies P1 and P2 are very close to the ideal case (around 2% difference in the worst case). Even without knowledge about the future (policy P2), energy gains are quite similar. This means that even simple shutdown policies – not including workload predictions – can save consequent amounts of energy, close to the optimal bound. These results show that the energy gains of P1 and P2 is too close (for Orion 0.28% of difference between the policies, roughly 2,000kWh over 6 years) to justify the elaboration of a prediction algorithm: such a complex algorithm to design would only bring negligible benefits.

6 Conclusion and Future Work

Energy consumption is more and more a worrying concern for Cloud data centers. Although shutdown techniques are available to reduce the overall energy consumption during idle periods, they are rarely employed because of their supposed impact on hardware.

Simulation results combining real workload traces and energy calibration measurements conducted in this paper allow us to draw several conclusions:

- Shutdown techniques can save important amounts of energy otherwise wasted during idle periods
- Even aggressive shutdown policies have no negative impact on disk lifetime.
- Reducing the consumption while in Off state has a greater impact on energy savings than reducing the switching energy and time costs between On and Off states. For this reason, S3 (Suspend-to-RAM) and S4 (Suspend-to-Disk) states are currently not beneficial in terms of energy consumption.
- Workload prediction is not worth the few energy it can save.

Our future work includes an integration of failure models when resuming from Off state in order to study the impact of bad resuming behavior. We also plan to evaluate other shutdown policies which are applied in current data centers like switching nodes by portions of the total number to control the impact on data center cooling system. We would like to explore heterogeneous architectures such

as ARM big.LITTLE for instance to see whether future architectures closer to energy-proportionality could still benefit from shutdown techniques.

Acknowledgments

Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see `https://www.grid5000.fr`).

This work is integrated and supported by the ELCI project, a French FSN ("Fond pour la Société Numérique") project that associates academic and industrial partners to design and provide software environment for very high performance computing.

References

1. D. Balouek, A. Carpen Amarie, G. Charrier, F. Desprez, E. Jeannot, E. Jeanvoine, A. Lèbre, D. Margery, N. Niclausse, L. Nussbaum, O. Richard, C. Pérez, F. Quesnel, C. Rohr, and L. Sarzyniec. Adding virtualization capabilities to the Grid'5000 testbed. In I. Ivanov, M. Sinderen, F. Leymann, and T. Shan, editors, *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 3–20. Springer International Publishing, 2013.
2. M. Daydé, B. Depardon, A. Franc, J.-F. Gibrat, R. Guilllier, Y. Karami, F. Suter, B. Taddese, M. Chabbert, and S. Thérond. E-Biothon: an experimental platform for BioInformatics. In *International Conference on Computer Science and Information Technologies (CSIT)*, pages 1–4, 2015.
3. A.-C. Orgerie, M. Dias de Assunção, and L. Lefèvre. A Survey on Techniques for Improving the Energy Efficiency of Large-scale Distributed Systems. *ACM Computing Survey*, 46(4):47:1–47:31, Mar. 2014.
4. A.-C. Orgerie and L. Lefèvre. ERIDIS: Energy-Efficient Reservation Infrastructure for Large-scale Distributed Systems. *Parallel Processing Letters*, 21(02):133–154, 2011.
5. A.-C. Orgerie, L. Lefèvre, and J.-P. Gelas. Save Watts in Your Grid: Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems. In *IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, pages 171–178, Dec 2008.
6. Seagate. Desktop HDD specification sheet. `http://www.seagate.com/staticfiles/docs/pdf/datasheet/disc/desktop-hdd-data-sheet-ds1770-1-1212us.pdf`, 2012.
7. Seagate. NAS HDD specification sheet. `http://www.seagate.com/www-content/product-content/nas-fam/nas-hdd/_shared/docs/nas-hdd-8tb-ds1789-5-1510DS1789-5-1510US-en_US.pdf`, 2015.
8. A. B. Yoo, M. A. Jette, and M. Grondona. *International Workshop Job Scheduling Strategies for Parallel Processing (JSSPP)*, chapter SLURM: Simple Linux Utility for Resource Management, pages 44–60. Springer, 2003.